Testing for evolutionary relationship: an application of mathematics in molecular biology

Barbour, A D

Abstract: The paper illustrates some mathematical problems that are motivated by molecular biological applications and the techniques that are used to address them. The context is that of detecting evolutionary relationship on the basis of molecular sequence data and of measuring its strength. The Stein–Chen method is shown to play a central role in the theoretical analysis of many of the procedures used in practice.

DOI: https://doi.org/10.1142/S0219607705000036

Testing for evolutionary relationship: an application of mathematics in molecular biology.

A. D. Barbour[1]
Universität Zürich

Abstract. The paper illustrates some mathematical problems that are motivated by molecular biological applications, and the techniques that are used to address them. The context is that of detecting evolutionary relationship on the basis of molecular sequence data, and of measuring its strength. The Stein–Chen method is shown to play a central rôle in the theoretical analysis of many of the procedures used in practice.

Keywords: Sequence alignment, tree construction, profile matching, DNA.

AMS 2000 Subject Classifications: 92D20, 92D15, 62E17.

[1] Institut für Mathematik, Winterthurerstrasse 190, CH–8057 ZÜRICH
Partially supported by Schweizerischer Nationalfondsprojekt 20–61753.00

1. Introduction

One of the scientific questions that has most consistently fascinated human beings is that of their own origins and place in the world; the chain of evolution which has led from the beginning of life on earth to present day man, and the degree of relationship between man and other animals. Evolution is understood to be a sequential process of small changes, adaptations in response to selection, and evolutionary descent can be established by finding a chain of intermediate species — most usually evidenced by their fossil relics — each of which is similar to the next, leading from some ancestor species to a contemporary descendant. The degree of relationship between two contemporary species can also be established in the same way, by noting where their chains of descent diverge. This method, however, involves an enormous investment of time and effort, as well as a fair degree of luck, since there is no guarantee that the fossil record will be well enough preserved to yield a chain without gaps; indeed, beyond a certain distance into the past, it is unlikely that such a chain could be uncovered.

A plausible alternative is to suppose that one can measure the similarity between pairs of species in some objective way. Degree of relationship is then assumed to be reflected in similarity, since more distantly related species, whose lines of descent diverged longer ago, have spent more time separately undergoing evolution, and hence can be expected to be less similar to one another. Starting from the matrix of pairwise similarities between a number of species, there are many algorithms (Waterman 1995, Ch. 14; references in §15.7, p. 385; Swofford *et al.*, 1996; Strimmer and von Haeseler, 1996) which produce family trees designed to be, in some way, most consistent with it; these trees can then be used as inferred evolutionary trees. There is clearly a whole field of mathematical investigation which lies behind such procedures, but this is not the goal of our paper. We shall concentrate simply on the problem of measuring similarity.

The classical approach is morphometric. Many different, clearly defined quantities — typically numerical measurements of dimensions, often also the presence or absence of features — are recorded for each of a number of representatives of a species. These are then combined into a characteristic profile for the species, perhaps consisting of the means, standard deviations and correlations derived from the numerical measurements, and the frequencies with which the features are present. Measuring similarity can then be reduced to measuring closeness (or its opposite, distance) in the high dimensional space of profiles, a problem which is, at least in principle, mathematically tractable. The measurements and

features involved are usually tangible, or at least microscopically visible, properties of the organism.

In more recent times, other measurements have come to replace dimensions and appearance. An important development was the discovery that useful profiles, characterizing closely related species or genetically distinct subgroups of a single species, could be derived by means of the size profiles of cellular enzymes, which can be directly compared using the technique of enzyme electrophoresis (Harris and Hopkinson, 1978). The enzymes are induced to travel along a gel by an electric field, the speed with which they do so — and hence their displacement over the time for which the experiment is conducted — being related to their size. The resulting displacement patterns can then be compared for similarity; for instance, in the framework above, having an enzyme of a particular size (or achieving a particular displacement) can be interpreted as possessing a particular feature.

Nowadays, objects even more microscopic have replaced these features: the amino acids at each position in the chains making up enzymes and proteins, and the letters forming the DNA strings which (in part) encode the blueprints for constructing them. The problem is still formally the same as before, but there are a number of important practical differences. One is that all the features are now discrete, since there are only four DNA letters, and only 20 amino acids; another is that there are usually hundreds or thousands of positions in the strings being compared (corresponding to profiles in a space of as many dimensions); another is that there is no absolute standard for determining which positions in the amino acid chains of DNA strings from two different species should correspond to one another; and, finally, the *order* in the sequence implies certain relationships between the features, in particular with regard to the actual, three dimensional structure of the molecules. These differences together result in data of a fundamentally different form, for which special techniques are needed.

2. Matching two sequences: empirical.

A first, natural way of measuring similarity between two strings of letters — from the DNA or amino acid alphabets — is to line them up, and to count how many positions there are in which the letters are the same in both sequences; this leads to the so called 'percentage identity' measure. The challenge in practice is to align them. For comparing DNA sequences from very closely related (sub–)species, such as different populations of human beings, or even between a human being and a chimpanzee, this may not be too difficult. The simplest tool is the 'dotplot', a matrix of 1's (dots) and 0's (spaces), where

a 1 at position (i, j) indicates that the letter at the i'th position of the first sequence is the same as that at the j'th position of the second sequence. If, in reality, position i in the first sequence corresponds to position j in the second, then, in a perfect alignment, position $i + k$ in the first corresponds to position $j + k$ in the second, for all integer k. For sequences that are almost identical, for instance the DNA sequences from two human beings, these pairs of positions will mostly give rise to 1's, and a diagonal of dots will manifest itself in the matrix, indicating the correct alignment.

There will in practice be occasional spaces, where the sequences have different letters at a given position — single nucleotide polymorphisms may lead to differences in the human genome at a rate of one difference in a few thousand letters — and the diagonal may occasionally 'shift', when one of the sequences has extra positions that are not present (gaps) in the other. The latter phenomenon frequently occurs in parts of the genome where short strings (e.g. triplets) are repeated many times, and natural copying errors lead to variation in the number of repeats. Such areas are of particular interest to forensic science, since they offer the best discrimination between individuals of a genetically relatively homogeneous population; some of them are also of significant medical importance, since, in certain areas, too many repeats can result in disease. But, for such comparisons, it is easy enough to devise a measure of dissimilarity, based on the number or proportion of differences at correctly matched positions and the number of shifts needed to maintain the correct alignment, with only the relative weight to be attached to these two forms of discrepancy as a subjective parameter, to be chosen by the scientist. Such comparisons lie behind statements such as 'humans and pygmy chimpanzees are genetically 98.4% identical'. The main difficulty lies in the fact that, for distinguishing degrees of dissimilarity between very similar individuals or populations, it may be necessary to examine very long sequences of letters, in order to find enough differences to give the required discriminatory power.

The longer the two species have evolved separately, the more difficult it becomes to distinguish the correct alignment from a dotplot. If mutations occur independently and at random at each position of a DNA sequence, then the letter appearing at a given position can be modelled over time as the value of a Markov chain, with the alphabet A,C,G,T as state space. The corresponding infinitesimal transition matrix may well be taken to exhibit some structure deriving from biological reality — for instance, G–C and A–T transitions being preferred — but, even so, the information about the distribution of the letter at a given site at time t contained in the letter that was there at time 0 decays exponentially with t, at a rate more or less depending on the overall mutation rate, ρ. If ρt is large, where $t/2$ denotes the time since the evolutionary chains of the two species diverged (each species is subject

to mutation), then the chance of the letters at correctly matched positions being the same is little different from the chance at totally different positions. In such circumstances, there is no hope of visually recognising the correct diagonal.

In order to increase the value of t for which similarities can be measured in this way, the computer can be used to evaluate all possible alignments (allowing shifts), and to find the one with the highest (log-)likelihood. This is essentially equivalent (because the DNA letters normally have more or less the same frequencies of 1/4) to finding the alignment with the highest score, when each match (position in the alignment where the letters of the two sequences are the same) contributes a fixed positive score, and each mismatch and gap different but fixed negative scores; basing consideration on likelihood implies particular values for these scores: see, in particular, Thorne, Kishino and Felsenstein (1991) and Durbin *et al.* (1998). Needleman and Wunsch (1970) introduced a dynamic programming algorithm which finds the optimal alignment relatively fast (in order $O(mn)$ steps, where m and n, not necessarily equal, are the lengths of the two sequences), so that this is a practical possibility for quite long pairs of sequences. Sequence similarity can then be related to the value of the optimal score obtained. However, in order to be able statistically to distinguish whether a proposed alignment is genuine or not, when the underlying Markov chain at each position approaches equilibrium at rate $O(e^{-\rho t})$, the two sequences must be at least of order $O(e^{\rho t})$ letters long. This exponential factor leads to practical limitations, and the global method can break down for values of t which are still relatively small on an evolutionary scale.

In order to be able to discriminate for still larger values of t, selection is invoked. The calculations above are based on the assumption that changes in the DNA sequence are allowed to occur at random. However, any such change which materially affects the well–being of the organism is subject to selection; frequently deleterious, occasionally advantageous. Thus changes in the DNA which codes for proteins and enzymes are strictly limited to changes that do not have a deleterious effect on their function; these changes can be expected to occur much more slowly, corresponding to values of ρ in the above model which are orders of magnitude smaller, and hence to values of t for which discrimination can be made which are orders of magnitude larger. As a result, if greater evolutionary distances are of interest, it is natural to compare the amino acid sequences of the enzymes and proteins themselves, or that part of the DNA which codes for them, rather than huge chunks of the original genome. Thus the underlying model has changed its parameters somewhat; from sequences of millions of letters from a four letter alphabet to sequences of hundreds of letters from a twenty letter alphabet.

Pushing this argument further still, the function of an enzyme is frequently determined by a particular 'fold' or 'active site', a local geometrical arrangement of amino acids which enables specific reactions to take place, and the remaining part of the enzyme merely keeps the fold stable. Since there may be many ways in which this remaining part may be able to achieve stability, but the fold itself may be extremely specifically determined, even greater values of t can be investigated, if similarity is restricted to (strong) similarity between substrings of the original sequences.

Two further considerations have thus been introduced, and need to be addressed. The first is that the concepts match and mismatch have to be refined. The DNA substitutions considered above were tacitly assumed to be selectively neutral, merely a (random) mark of the passage of time, and hence equivalent. In contrast, for proteins evolving under selection, some amino acid substitutions may be generally of little importance, others much more significant, because of similarity or dissimilarity with respect to size, electrical charge, hydrophobicity and so on. In a pioneering paper, Dayhoff *et al.* (1978) introduced the empirical, likelihood based 'PAM' family of scoring matrices, designed to give weights to particular mismatches, appropriate to the pair of amino acids involved and to the evolutionary distance at which comparison is being made; current alternatives include the BLOSUM family (Henikoff and Henikoff, 1992) and the more sophisticated approach of Müller and Vingron (2001), again based on likelihood considerations, in which the typical evolutionary distance between the species under consideration need not be specified in advance.

The second consideration is that the optimal alignment now refers not to a global match of two amino acid strings, but to the best match of a pair of suitably chosen substrings, still perhaps allowing gaps. Finding the optimal substring alignment for a given mismatch scoring matrix and gap penalty turns out to be possible by modifying the Needleman and Wunsch algorithm (Smith and Waterman, 1981), with running time essentially unchanged, and since the strings involved are now much shorter, the computational effort (at least for a single pair of proteins) is now feasible: for whole database searches, faster heuristics such as FastA (Pearson and Lipman, 1988) or BLAST (Altschul *et al.*, 1990) are still perhaps to be preferred. But how does one distinguish whether the resulting score reflects real similarity, or just a best value that has been thrown up by chance? This problem, even when simplified to well specified mathematical models, is a difficult one. In the next section, we illustrate how serious mathematics is motivated by problems of molecular biology, by discussing it in more detail.

3. Matching two sequences: mathematical.

We start with almost the simplest abstraction of the situation above. We let $\xi_1, \ldots, \xi_m$ and $\eta_1, \ldots, \eta_n$ be two independent sequences of independently chosen letters from a finite alphabet $\mathcal{A}$, the ξ_i chosen according to a distribution μ and the η_j according to ν. Fix k, and set

$$I_{ij} := I[\xi_i = \eta_j, \xi_{i+1} = \eta_{j+1}, \ldots, \xi_{i+k-1} = \eta_{j+k-1}], \tag{3.1}$$

so that $I_{ij} = 1$ means that there is a perfectly matching substring alignment of length k which begins at position i of the first sequence and at position j of the second. Thus the event that

$$W := \sum_{i=1}^{m-k+1} \sum_{j=1}^{n-k+1} I_{ij} > 0 \tag{3.2}$$

means that there is at least one perfectly matching substring alignment of length k, somewhere in the sequences. Note the simplifying model assumptions: perfect matching is required, no gaps are allowed in either substring, and different positions have independently assigned letters (amino acids). The idea is to compare the length k of the longest such substring match obtained between a pair of actual amino acid or DNA sequences with the probability that a substring match of that length or longer could have occurred by chance in the above, idealized model, and to use this as a measure of its 'significance'. Hence primary mathematical interest centres on the probability $\mathbb{P}[W > 0]$ for different values of k.

The random variable W is a sum of identically distributed 0–1 random variables I_{ij} which, for interesting values of k, have small probability p^k of taking the value 1, where

$$p := \sum_{a \in \mathcal{A}} \mu\{a\}\nu\{a\}; \tag{3.3}$$

furthermore, I_{ij} is independent of $\{I_{rs} : |i - r| \geq k \text{ and } |j - s| \geq k\}$. This suggests that a Poisson approximation to the distribution of W may be reasonable, based on the following remarkable theorem of Chen (1975), now widely used under the title 'Stein–Chen' method.

Suppose that a random variable W can be expressed as a sum $\sum_{\gamma \in \Gamma} I_\gamma$, where Γ is any finite index set and the I_γ are 0–1 random variables. For each $\gamma \in \Gamma$, suppose that N_γ is a subset of Γ containing γ, and define

$$X_\gamma := \sum_{\beta \in N_\gamma \setminus \{\gamma\}} I_\beta; \quad W_\gamma := \sum_{\beta \notin N_\gamma} I_\beta. \tag{3.4}$$

Then set

$$\varepsilon_0 := \sum_{\gamma\in\Gamma} \mathbb{E}|\mathbb{P}[I_\gamma = 1 \,|\, W_\gamma] - \mathbb{P}[I_\gamma = 1]|; \quad \varepsilon_1 := \sum_{\gamma\in\Gamma} \{\mathbb{E}(I_\gamma X_\gamma) + \mathbb{E}I_\gamma(\mathbb{E}X_\gamma + \mathbb{E}I_\gamma)\}. \tag{3.5}$$

Then, *whatever the choices of subsets* N_γ, the total variation distance between the distribution of the random variable W and the Poisson distribution with mean $\lambda := \mathbb{E}W$ is no larger than

$$\min\{1, \lambda^{-1/2}\}\varepsilon_0 + \min\{1, \lambda^{-1}\}\varepsilon_1. \tag{3.6}$$

The random variable W defined in (3.2) is a sum of 0–1 random variables I_{ij} with finite index set $\Gamma := \{(i,j) : 1 \le i \le m, 1 \le j \le n\}$, and $\mathbb{P}[I_{ij} = 1] = p^k$ is small for 'interesting' values of k, which are those when $\lambda = mnp^k$ is small (implying in particular that $\mathbb{P}[W > 0]$ is small). Thus, for instance, the final element $\sum_{\gamma\in\Gamma}\{\mathbb{E}I_\gamma\}^2$ in ε_1 takes the value $p^k\lambda$, which is very much smaller even than λ. Choosing

$$N_{ij} := \{(r,s) : |i-r| < k \text{ or } |j-s| < k\}$$

to exploit the independence structure, it follows that I_{ij} is independent of W_{ij}, so that $\varepsilon_0 = 0$. Then

$$\mathbb{E}X_{ij} \le (2k-1)(m+n)p^k,$$

so that the second element $\sum_{\gamma\in\Gamma} \mathbb{E}I_\gamma \mathbb{E}X_\gamma$ in ε_1 is itself at most

$$(2k-1)(m+n)p^k\lambda = \{(2k-1)(m+n)/mn\}\lambda^2,$$

which is a small fraction of the very small λ^2 whenever, as is typically the case in practice, $k \ll \min\{m,n\}$. This all looks very good for Poisson approximation, but the remaining element $\sum_{\gamma\in\Gamma} \mathbb{E}(I_\gamma X_\gamma)$ in ε_1 is not so small: if $I_{ij} = 1$, the conditional probability that then $I_{i+r,j+r} = 1$ is p^r, $1 \le r \le k-1$, which may be quite large when r is small (for $r = 1$, about $1/4$ for DNA sequences, and about $1/20$ for amino acid sequences), even though p^k is small. In particular, using just $r = 1$, it follows that

$$\mathbb{E}(I_{ij}X_{ij}) \ge p\mathbb{E}I_{ij},$$

and hence that

$$\sum_{\gamma\in\Gamma} \mathbb{E}(I_\gamma X_\gamma) \ge p\lambda.$$

Thus this term in the error bound given in (3.6) is of about the same size as the probability $(1 - e^{-\lambda})$ that would result from the Poisson approximation to $\mathbb{P}[W > 0]$, which is at the very least mathematically unsatisfying. What is more, the apparent inaccuracy is no artefact of the method; there is some local clustering of 1's among the I_{ij}, and a compound Poisson approximation to the distribution of W is in fact more appropriate. Hence the attempt to approximate the probability $\mathbb{P}[W > 0]$ by using Poisson approximation appears to have failed.

Fortunately, the problem can be circumvented, following the approach taken in Arratia, Goldstein and Gordon (1989). Local clustering is avoided if the I_{ij} are replaced by related random variables I'_{ij}, where

$$I'_{ij} := I[\xi_{i-1} \neq \eta_{j-1}]\, I_{ij};$$

then $W' := \sum_{i=1}^{m-k+1} \sum_{j=1}^{n-k+1} I'_{ij} \neq W$ in general, but (neglecting edge effects) the events $W > 0$ and $W' > 0$ are identical, so that the required probability can be deduced from approximation to the distribution of W' instead. The indicator random variables I'_{ij} are, much as before, close to being independent, with I'_{ij} independent of $\{I_{rs} : |i - r| > k \text{ and } |j - s| > k\}$, but those that are strongly dependent are now *negatively* correlated, so that there is little tendency for there to be clusters. Taking

$$N_{ij} := \{(r, s) : |i - r| \leq k \text{ or } |j - s| \leq k\},$$

one still has $\varepsilon_0 = 0$, and the bound $\mathbb{E}X_{ij} \leq (2k+1)(m+n)p^k$ is almost unchanged; but now $\mathbb{E}(I_{ij} I_{i+r,j+r}) = 0$ for $1 \leq r \leq k$, so that the troublesome contributions to $\sum_{\gamma \in \Gamma} \mathbb{E}(I_\gamma X_\gamma)$ have disappeared, and the remaining elements are of manageable size. Thus the Stein–Chen method applied to the random variable W' shows that an approximation by the Poisson distribution with mean $\mathbb{E}W' = mnp^k(1-p)$ is accurate up to the error given in (3.6), which is now typically small.

In particular, if the distributions μ and ν are both uniform over the letters of the alphabet $\mathcal{A}$, the bound in (3.6) is in all at most of order $O(k(m + n)mnp^{2k})$. Hence the error in the approximation

$$\mathbb{P}[W' > 0] \approx 1 - \exp\{-\mathbb{E}W'\} = 1 - \exp\{-mnp^k(1 - p)\}$$

is at most of *relative* order $O(k(m+n)p^k)$, which for the interesting values of k is typically like $O(\lambda m^{-1} \log n)$ if $n \geq m \gg 1$. This is at once an elegant mathematical theorem and a practically useful estimate.

There are two directions in which the theorem is oversimplified, from the point of view of practical application. The first is that gaps are not allowed. This problem was addressed in Neuhauser (1994). Here, the basic element is no longer a matching substring of length k, but a matching substring of length k allowing for gaps of total length at most t: the previous problem had $t = 0$. The index set for the indicators of such events is immediately very much larger, since now the starting points of all contiguous segments have to be recorded, and the adjustment to the indicators to avoid the occurrence of clusters, akin to passing from W to W' above, becomes more complicated. Indeed, Neuhauser actually counts the number W'' of suitable gapped substrings found by a particular greedy algorithm, a number which is easier to handle, and for which the probability $\mathbb{P}[W'' = 0]$ is asymptotically equivalent to the probability required. Nonetheless, just as in the simpler problem, the indicators only exhibit local dependence, and the distribution of W'' can be well approximated by the Poisson distribution with the same mean, again with an error which is explicitly bounded by applying the theorem of Chen (1975).

The second oversimplification is the consideration only of exact matching, which is generally too restrictive to be used with amino acid sequences. Very general scoring schemes have been studied by Dembo, Karlin and Zeitouni (1994), in the case of contiguous substrings. Once again, the independence structure makes Poisson approximation by way of the Stein–Chen method the appropriate tool, as soon as the effect of local clustering has been eliminated. To do this involves more sophisticated arguments based on the properties of general random walks, together with a choice of scoring scheme which ensures that the length of the optimal substring match (the analogue of k above) is much smaller than the lengths of the sequences being compared; a minimal requirement is that the expected score from a randomly matched pair of positions should be negative. Their approach leads to good approximations for the probability on the random model of obtaining a substring match with a prespecified large score. Thus theoretical significance values can be associated with any choice of critical score threshold. However, their results are not applicable to substrings in which gaps are allowed.

In practice, gaps are a necessary feature of alignments, and such approximations need to be extended to cover this setting also. This has proved to be an awkward task. There is a recent heuristic of Mott and Tribe (1999), which seems to work well in practice, and an even more recent large deviation theorem of Siegmund and Yakir (2000), which leads to similar approximate p–values, but which is only proved in an asymptotic setting which

includes the (inconvenient) assumption that the penalty for opening a gap becomes large in the limit. Thus, even today, the combination of general scoring schemes with gaps has still not been treated to full mathematical satisfaction.

4. Sequence to profile matching.

Using the Smith–Waterman (1981) algorithm, it is possible to take any new protein sequence, and to compare it successively with every element of a database of known proteins, such as SWISSPROT, in order to look for possible relatives. This procedure raises further problems, both practical and mathematical. The previous analyses have been based on computing ideal significance probabilities for the comparison of a single pair of sequences. If many comparisons are to be made simultaneously, the significance probabilities for the individual comparisons must be made very small, if the chance of obtaining at least one false positive is to be kept small. SWISSPROT contains of the order of 100'000 well documented proteins, and so to keep the chance of a false positive to 5%, when comparing a protein with the whole database, the significance levels for the individual comparisons have to be of the order of $(5/100'000)$ % (maybe not quite so small, because the database is far from being a random sample of proteins, but still very small). If the individual significance levels are kept so low, the search is unlikely to be able to distinguish any distant relatives; if not, there will be many false positives to be eliminated by other means. Thus, if distant relatives are to be discovered, more information is needed.

One technique which has proved very effective is to use the additional information contained in the family of known relatives of a protein. If a number of proteins are known to be related — on biochemical grounds, or because they are individually strongly related on the basis of pairwise sequence comparison — the aim is to simultaneously combine the information in the individual sequences to produce a 'multiple' alignment. This consists of a sequence of positions, along which each sequence is arranged, with gaps if necessary. At each position, there is information from each of the sequences, in the form of the amino acids (now including 'gap') represented there. Positions at which the empirical distribution of these amino acids is highly concentrated ('conserved' positions) are interpreted as carrying important biochemical information, and are given a high loading in subsequent comparisons.

The resulting sequence of empirical distributions is called a 'profile' for the family (Gribskov *et al.*, 1990). Formally, if the multiple alignment of the n sequences is of length l and the value (amino acid residue or gap) of the k'th sequence at the i'th position in the alignment is denoted by a_{ki}, then an empirical probability distribution P_i for the value at

position i is defined by

$$P_i\{\alpha_j\} := \sum_{k=1}^{n} w_k 1_{\{a_{ki}=\alpha_j\}}, \quad 1 \le j \le 21, \tag{4.1}$$

where $w_1, \ldots, w_n$ denote the weights attached to the sequences of the family (see below), chosen so that $\sum_{k=1}^{n} w_k = 1$, and $\alpha_1, \ldots, \alpha_{20}$ denote the 20 amino acids, α_{21} the value 'gap'. The profile for the family then consists of the sequence $\mathcal{P} := (P_1, P_2, \ldots, P_l)$. Any new sequence $A' := (a'_i : 1 \le i \le m)$, can be scored for similarity to the family by computing its best alignment h with the profile, usually by maximizing an additive score

$$T(\mathcal{P}, A'; h) := \sum_{i=1}^{L(h)} W(P_{h(1,i)}) S(P_{h(1,i)}, a'_{h(2,i)}) \tag{4.2}$$

over alignment functions $h := (h(1,\cdot), h(2,\cdot))$. These functions satisfy

$$h(1,i) \in \{0,1,\ldots,l\}; \quad h(2,i) \in \{0,1,\ldots,m\}; \qquad 1 \le i \le L(h),$$

for some $L(h)$ such that $\max\{l,m\} \le L(h) \le l+m$, and they determine an alignment in the sense that $h(1,\cdot)$ ($h(2,\cdot)$) takes each of the values $1,2,\ldots,l$ ($1,2,\ldots,m$) exactly once and in increasing order, and $h(1,i) = h(2,i) = 0$ is not allowed. For the remaining elements appearing in (4.2), $W(P)$ denotes the loading associated with a comparison at a position whose profile distribution is P, and is usually a function its concentration; and $S(P,a)$ is a score function representing the plausibility of finding value a at a position where the empirical distribution of amino acids is P; and finally, P_0 and a'_0 are both interpreted as 'gap'. If a new sequence has a significantly high score when compared to the profile, it may be taken to be a relative of the family.

The advantage of a profile is that it emphasizes the essential features common to the members of the family, through the loadings $W(P)$; in this way, subsequent sequence matching to the profile concentrates on those parts of the sequence which are biologically relevant, and reduces the likelihood of false positives arising as a result of chance substring matches in some irrelevant part of the sequence. Practical experience shows that comparing new sequences to family profiles substantially enhances the efficacy of detection of distant relatives.

There are a number of algorithms for arriving at a multiple alignment (Waterman 1995, Ch. 10; references in §15.4.3, p. 382). These are often based on sequential merging

procedures, in which, at each stage, a best alignment is found between the next introduced sequence and the current multiple alignment; another approach is by way of hidden Markov models (Krogh *et al.* 1994, Neuwald *et al.* 1997). One major problem is how to weight the information in the different sequences (Altschul, Carroll and Lipman, 1989; Thompson, Higgins and Gibson, 1994), since two (almost) identical sequences are frequently repeating the same item of evolutionary information, having diverged only very recently, whereas two more widely differing sequences could be expected to have only the biologically essential information in common, their comparison thus being potentially much more informative. A reasonable strategy is thus first to construct a rough family tree based on the pairwise sequence similarities, and then to use this to determine weights for the sequences in the multiple alignment procedure. However, once a multiple alignment has been constructed, it includes in particular the 'preferred' alignment of each individual pair of sequences. This can in turn be used, together with the positional weights, to determine new pairwise similarities and hence to redefine the weights for the sequences, thereby leading to a new multiple alignment; this whole procedure can then be iterated until convergence is reached. Other more refined techniques have been considered: see, for example, Vingron and von Haeseler (1997).

Assigning an objective measure of quality to a multiple alignment is thus a challenging task. So, too, is the assessment of the significance of the score resulting from the comparison of a new sequence to the profile. A method commonly used in practice is to compare a large number of unrelated sequences with the profile, and to fit an extreme value distribution to the upper tail of the empirical distribution of the scores thus obtained; this distribution is then used to assess the significance of further scores.

5. Profile to profile matching.

The practical success of sequence to profile matching in finding distant relatives suggests that it may be of further help to use family information for the 'new' sequences as well. This is at first sight impossible, since a new sequence has no known relatives. However, an 'empirical' family can be constructed by applying sequence comparison routines, without reference to pre–existing biological knowledge. The new sequence can be compared to each member of the database by quick procedures such as FastA or BLAST, which look for high scoring local substring matches. From those sequences selected as having very high scores, a profile can be constructed. Now the profile can be compared with all members of the database, new relatives being incorporated, and badly fitting members from the initial

search rejected. This determines the empirical family associated with the new sequence. At worst, if the sequence shows no similarity at all to any sequence currently in the database, the empirical family will consist merely of the original sequence; otherwise, the procedure picks out some aspect of the sequence which is mirrored elsewhere in the database, and the other sequences which reflect it.

In the FPA procedure of Mehta *et al.* (1999), the new sequence is tested for relationship with a known, biochemically determined family by comparing each member of its empirical family with the profile of the known family, using sequence to profile comparison. The average of these profile scores (weighted to reflect the relationships between the sequences in the empirical family, in the same way that is used when constructing a profile) is then taken as the overall score for the comparison. A reference distribution is determined empirically, by drawing 100 sequences at random from the database, constructing their 100 empirical families in the same fashion as for a new sequence, and then computing the corresponding 100 average profile scores. This gives an empirical null distribution, with which the actual score obtained can be compared. Since this null distribution typically resembles a normal distribution, because of the averaging, its median and median absolute deviation are used as a basis for determining 'significant' scores (at least 5 MAD's larger than the median).

FPA has proved useful in practice in suggesting new relationships between proteins (Mehta *et al.*, 1999), and it appears to extend the evolutionary distance at which relationship can still be determined a little further than sequence to profile comparison. However, the method used to compare the two families, the empirical and the known family, is asymmetric; a profile from the known family is compared sequence by sequence with the members of the empirical family. This has the advantage that existing, well documented software can be directly used. However, it may also be of interest to be able to compare two profiles directly, in a symmetric fashion, for instance if similarities between protein families or subfamilies are to be used to reconstruct the evolutionary history of these families. The following proposed procedure is derived from Siegrist (1998); an alternative procedure was suggested by Pietrokowski (1996), but the underlying distance measure employed there seems unsuited to the context.

In order to align two profiles $\mathcal{P} = (P_1, P_2, \ldots, P_l)$ and $\mathcal{Q} = (Q_1, \ldots, Q_m)$, we adopt the previous method of maximizing an additive score

$$T(\mathcal{P}, \mathcal{Q}; h) := \sum_{i=1}^{L(h)} T'(P_{h(1,i)}, Q_{h(2,i)}) \tag{5.1}$$

over alignment functions $h := (h(1,\cdot), h(2,\cdot))$ as in the previous section. The pair score

$T'(P_r, Q_s)$ is computed as

$$T'(P,Q) = W(P,Q)S(P,Q), \tag{5.2}$$

the product of a weight $W(P,Q)$, reflecting the 'relevance' of the distributions P and Q in their alignments, and a similarity $S(P,Q)$ between the two distributions. The final score is then computed as the weighted average of the similarities at the optimal alignment h^*,

$$T^*(\mathcal{P},\mathcal{Q}) := T(\mathcal{P},\mathcal{Q};h^*) \Big/ \sum_{i=1}^{L(h^*)} W(P_{h^*(1,i)}, Q_{h^*(2,i)}), \tag{5.3}$$

and thus always takes a value within the range of possible similarities $S(P,Q)$ between two distributions P and Q. This is a useful property for the interpretation of the results. It is, for instance, independent of the length of the common structural motif which exhibits the relationship between two related profiles. Further information, as yet unexploited, is to be found in the total weight $\sum_{i=1}^{L(h^*)} W(P_{h^*(1,i)}, Q_{h^*(2,i)})$ at the optimum.

The optimal alignment h^* is determined using a typical dynamic programming recursion. For $1 \le i \le l$, $1 \le j \le m$, define T_{ij} to be the optimal additive score, defined as in (2), for any alignment of $(P_1, \ldots, P_i)$ with $(Q_1, \ldots, Q_j)$, E_{ij} the optimum among all such alignments having Q_j matched with G, and F_{ij} the optimum among those having P_i matched with G. Then the E_{ij}, F_{ij} and T_{ij} are determined recursively; the quantities for given i, j are found from those with one or both indices smaller by setting

$$\begin{aligned} E_{ij} &= \max\{T_{i,j-1} + gS(G,Q_j), E_{i,j-1} + egS(G,Q_j)\}, \quad i \ge 0,\ j \ge 1; \\ F_{ij} &= \max\{T_{i-1,j} + gS(P_i,G), F_{i-1,j} + egS(P_i,G)\}, \quad i \ge 1,\ j \ge 0, \end{aligned} \tag{5.4}$$

and

$$T_{ij} = \max\{T_{i-1,j-1} + T'(P_i,Q_j), E_{ij}, F_{ij}\}, \quad i,j \ge 1. \tag{5.5}$$

The quantity g denotes the weight attached to an opening match of P or Q with G, and the factor e is used to reduce the weight if a gap is *extended* ($S(G,Q)$ and $S(P,G)$ are never positive): see (5.11). Appropriate initial values are obtained by taking $T_{i0} = F_{i0}$ and $T_{0j} = E_{0j}$, and starting the recursion (5.4) with $E_{00} = F_{00} = 0$. The alignment h^* which attains the optimum $T(\mathcal{P},\mathcal{Q};h)$ is then recovered by noting, at each step (i,j) of the recursion, which transition led to the maximal value T_{ij}, and then reconstructing backwards from T_{lm}: see Waterman (1995, Section 9.8). All that remains to determine our procedure is thus the specification of $S(P,Q)$ and $W(P,Q)$ in (5.2).

The similarity $S(P,Q)$ between two distributions P and Q which are concentrated on the values $\alpha_1,\ldots,\alpha_{20}$ we take to be

$$S(P,Q) = S(Q,P) := \max_{x\in X(P,Q)} \left\{ \sum_{i=1}^{20}\sum_{j=1}^{20} \Omega_{ij}x_{ij} \right\}, \tag{5.6}$$

where $X(P,Q)$ denotes the set of all non–negative 20×20 matrices satisfying

$$\sum_{k=1}^{20} x_{ik} = P\{\alpha_i\} \text{ and } \sum_{k=1}^{20} x_{ki} = Q\{\alpha_i\}, \quad 1\le i\le 20,$$

and Ω is a matrix of pairwise similarity scores. The matrix Ω should, as at the end of Section 3, be chosen in such a way that the average similarity of two randomly chosen amino acids is sufficiently negative that long high scoring substring matches are very unlikely to occur at random. The measure of similarity defined in (5.6) is related to distances of Monge–Kantorovich type (Rachev 1991, Chapter 5), and it can rapidly be computed using the Ford–Fulkerson transportation algorithm. The optimal matrix x determines that probability distribution over *pairs* of amino acids which maximizes the average Ω–similarity of a randomly chosen pair of amino acids, subject to the requirement that x should have marginals P and Q. If either of P or Q puts probability on α_{21} (gap), we define $S(P,Q) = S(P',Q')$, where

$$P'\{\alpha_j\} = P\{\alpha_j\}/(1-P\{\alpha_{21}\}), \quad Q'\{\alpha_j\} = Q\{\alpha_j\}/(1-Q\{\alpha_{21}\}), \quad 1\le j\le 20; \tag{5.7}$$

P' and Q' are the conditional distributions of P and Q, given that gap does not appear. The similarity S is measured on the same scale as the matrix Ω measures similarity between pairs of amino acids, making it easily interpretable; the value of $T^*(\mathcal{P},\mathcal{Q})$ is thus also interpretable on the same scale. In particular,

$$S(P,Q) \le \Omega_* \quad \text{for all } P,Q,$$

where Ω_* is the common value of Ω_{ii}, $1\le i\le 20$, and $S(P,P)=\Omega_*$ for all P.

The effect of the gap probabilities $P\{\alpha_{21}\}$ and $Q\{\alpha_{21}\}$ is not felt in $S(P,Q)$, but is expressed instead in the weight $W(P,Q)$, which is defined to be

$$W(P,Q) := \{g(P)g(Q)\}^r \tag{5.8}$$

for some $r \geq 0$. The quantity $g(P)$ measures the concentration of the distribution P, again in terms of Ω–similarity, in that

$$g(P) := (1 - P\{\alpha_{21}\}) \sum_{i=1}^{20} \sum_{j=1}^{20} \Omega_{ij} P'\{\alpha_i\} P'\{\alpha_j\}. \tag{5.9}$$

The definition of $g(P)$ differs from that of $S(P, P)$ in (5.6), not only in the factor $(1 - P\{\alpha_{21}\})$, but also because there is no maximization with respect to x, in contrast to that in (5.6), which leads to $S(P, P) = \Omega_*$. Instead, x_{ij} is fixed as $P'\{\alpha_i\}P'\{\alpha_j\}$, and the double sum in (5.9) represents the average Ω–similarity between two amino acids chosen *independently* from the distribution P'. This average similarity is small if the distribution P' is very spread out, or if P' assigns probability to amino acids which have low Ω–similarity.

When comparing a distribution P with the gap G, we recognize similarity only through the value of $P\{\alpha_{21}\}$; we take

$$S(P, G) = S(G, P) := c + dP\{\alpha_{21}\}, \tag{5.10}$$

where $c \leq \min_{i,j} \Omega_{ij}$, the least possible similarity between distributions. The value c is obtained when $P\{\alpha_{21}\} = 0$, and $0 < d \leq -c$ increases the similarity between P and G in proportion to $P\{\alpha_{21}\}$, but not sufficiently to result in a positive similarity between any P and G.

The weight $W(P, G) = W(G, P)$ assigned to an alignment of P with G is not actually a function of P and G alone. Instead, the usual ideas of gap opening and gap extension penalties are adopted (see (5.4)); if $h(2, i) = 0$, then

$$W(P_{h(1,i)}, G) = \begin{cases} g & \text{if } h(2, i-1) \neq 0; \\ eg & \text{if } h(2, i-1) = 0, \end{cases} \tag{5.11}$$

where the gap opening weight g is taken to be Ω_*^{2r}, the largest weight that could possibly be assigned in (5.8) — this is then multiplied by the *negative* similarity in (5.10) — and the extension factor e, introduced in (5.4), reduces the penalty for extending a gap: often, one chooses $e = 0$.

The definitions (5.6)–(5.11) are used to determine the pair score $T'(P, Q)$ defined in (5.2), and the recursions (5.4) and (5.5) are used to maximize $T(\mathcal{P}, \mathcal{Q}; h)$ as in (5.1) with respect to alignments h; the optimal alignment h^* is then substituted into (5.3) to obtain the score $T^*(\mathcal{P}, \mathcal{Q})$ for the similarity of the two profiles $\mathcal{P}$ and $\mathcal{Q}$.

The procedure is still in an experimental stage, and further study, both theoretical and empirical, is required for a proper evaluation of the precise strength of evidence from such comparisons. Calibrating the significance of the scores obtained depends on the distribution of $T^*(\mathcal{P}, \mathcal{Q})$ for unrelated families, which are not so easy to model or to simulate as the 'empirical families' of FPA. Since there are many fewer families currently known than there are protein sequences, empirical determination of such extreme significance levels as are current for sequence to profile search, by using data from the profiles of actual protein families, cannot be achieved; on the other hand, by the same token, less extreme values are needed when fewer comparisons are made, for the same number of false positives.

References.

[1] S. F. Altschul, J. R. Carroll and D. J. Lipman (1989) Weights for data related by a tree. *Journal of Molecular Biology*, **207**, 647–653.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

[3] R. Arratia, L. Goldstein and L. Gordon (1989) Two moments suffice for Poisson approximations: the Chen–Stein method. *Annals of Probability*, **17**, 9–25.

[4] L. H. Y. Chen (1975) Poisson approximation for dependent trials. *Annals of Probability*, **3**, 534–545.

[5] M. O. Dayhoff, R. M. Schwarz and B. C. Orcutt (1978) Matrices for detecting distant relationships. *Atlas of Protein Sequences and Structure* **5**, 353–358.

[6] A. Dembo, S. Karlin and O. Zeitouni (1994) Limit distribution of maximal non–aligned two–sequence segmental score. *Annals of Probability* **22**, 2022–2039.

[7] J. Devereux, P. Haeberli and O. Smithies (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids research* **12**, 387–395.

[8] R. Durbin, S. Eddy, A. Krogh and G. Michison (1998) *Biological sequence analysis.* Cambridge University Press.

[9] M. Gribskov, R. Lüthy and D. Eisenberg (1990) Profile analysis. *Methods in Enzymology* **183**, 146–159.

[10] H. Harris and D. A. Hopkinson (1978) *Handbook of enzyme electrophoresis in human genetics.* North Holland, Amsterdam.

[11] S. Henikoff and J. G. Henikoff (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Science* **89**, 10915–10919.

[12] A. Krogh, M. Brown, I. S. Mian, K. Sjölander and D. Haussler (1994) Hidden Markov models in computational biology: Applications to protein modelling. *Journal of Molecular Biology*, **235**, 1501–1531.

[13] P. K. Mehta, P. Argos, A. D. Barbour and Ph. Christen (1999) Recognizing very distant sequence relationships among proteins by family profile analysis. *Proteins* **35**, 387–400.

[14] R. Mott and R. Tribe (1999) Approximate statistics of gapped alignments. *Journal of Computational Biology* **6**, 91–112.

[15] T. Müller and M. Vingron (2001) Modeling amino acid replacement. *Journal of Computational Biology* (to appear).

[16] S. B. Needleman and C. D. Wunsch (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology* **48**, 443–453.

[17] C. Neuhauser (1994) A Poisson approximation for sequence comparisons with insertions and deletions. *Annals of Statistics* **22**, 1603–1629.

[18] A. F. Neuwald, J. S. Liu, D. J. Lipman and C. E. Lawrence (1997) Extracting protein alignment models from the sequence database. *Nucleic Acids Research* **25**, 1665–1677.

[19] W. R. Pearson and D. J. Lipman (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Science* **85**, 2444–2448.

[20] S. Pietrokovski (1996) Searching databases of conserved sequence regions by aligning protein multiple–alignments. *Nucleic Acids Research* **24**, 3836–3845.

[21] S. T. Rachev (1991) *Probability metrics and the stability of stochastic models.* Wiley, New York.

[22] P. R. Sibbald and P. Argos (1990) Weighting aligned proteins or nucleic acid sequences to correct for unequal representation. *Journal of Molecular Biology* **216**, 813–818.

[23] D. Siegmund and B. Yakir (2000) Approximate p–values for local sequence alignments. *Annals of Statistics* **28**, 657–680.

[24] S. Siegrist (1998) Ein symmetrischer Test auf Verwandtschaft zwischen Familien von Sequenzen. Diplomarbeit, Institut für Mathematik, Universität Zürich.

[25] T. F. Smith and M. S. Waterman (1981) The identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195–197.

[26] K. Strimmer and A. von Haeseler (1996) Quartet puzzling: a quartet maximum–likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution* **13**, 964–969.

[27] D. L. Swofford, G. J. Olsen, P. J. Waddell and D. M. Hillis (1996) Phylogenetic inference. In: *Molecular systematics*, Eds: D. M. Hillis, C. Moritz and B. K. Mable, pp. 407–514. Sinauer Associates, Sunderland.

[28] J. D. Thompson, D. G. Higgins and T. J. Gibson (1994) Clustal W: Improving the sensitivity of progressive multiple alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673–4680.

[29] J. L. Thorne, H. Kishino and J. Felsenstein (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution* **33**, 114–124.

[30] M. Vingron and A. von Haeseler (1997) Towards integration of multiple alignment and phylogenetic tree construction. *Journal of Computational Biology* **4**, 23–34.

[31] M. S. Waterman (1995) *Introduction to computational Biology.* Chapman and Hall, London.